Content based access for a massive database of human observation video

Laurent Joyeux, Erika Doyle, Hugh Denman, Andy Crawsord, Adrien Bousseau, Anil Kokaram, Ray Fuller

Content Based Access for a Massive Database of Human Observation Video

L. Joyeux[†], E. Doyle[‡], H. Denman[†], A.C. Crawford[†], A. Bousseau[†], A. Kokaram[†], R. Fuller[‡]
[†] Dept. of Electronic & Electrical Engineering
[‡] Dept. of Psychology
Trinity College Dublin, Ireland
e-mail:ljoyeux@mee.tcd.ie, edoyle4@tcd.ie, akokaram@tcd.ie, rfuller@tcd.ie

ABSTRACT

We present in this paper a CBIR system for use in a psychological study of the relationship between human movement and Dyslexia. The system allows access to up to 500 hours of video and is an example of a scientific user context. This user context requires 100% accurate indexing and retrieval for a set of specific queries. This paper presents a novel use of interactive visual and audio cues for attaining this level of indexing performance. Furthermore, the issue of motion estimation accuracy in the presence of compression artifacts is explored as part of the data integrity storage problem. In addition, content based motion analysis techniques accurate enough to parse sequences on the basis of motion and objectively evaluate that motion are investigated. The tool allows Psychologists to undertake a study that would previously be impractical and the paper presents a number of lessons gained from the ongoing work.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Measurement, Human Factors

Keywords

content retrieval, tracking, video retrieval, dyslexia, human body motion

1. INTRODUCTION

Developmental dyslexia (also known as 'Specific Learning Difficulty' or SLD) is a serious societal problem. It affects 8% of the population - that implies 480,000 people in Ireland alone. It is not caused by lack of intelligence, emotional disturbance, poor teaching, family difficulties or social problems. If left untreated, a child can develop poor self-esteem and confidence and fail to master even the basics of reading, writing and arithmetic. These children require a high level of educational resources and have the strong potential to continue causing problems in the school system. The cost of dyslexia to the society infrastructure as a whole is therefore enormous.

There is currently no reliable diagnosis available to identify dyslexia until the child has demonstrated a failure to read after persistent attempts (usually at the age of 8 or 9). Remedial therapies are based on intensive practice of basic language skills and so occupy a large amount of teacher resources (often on a one to one basis). More often than not the child never reaches his or her appropriate reading age.

McPhillips et al. [11] presented the notion that there is a quantifiable connection between Dyslexia and the retention of certain reflex movements. Dyslexia is now no longer seen solely as a problem generated by a higher-order brain malfunction, but as possibly a treatable disorder with a physiological rationale. Evidence was provided that in Dyslexics, certain *primary reflexes* [9] are retained. In subsequent development, these reflexes become integrated into postural reflexes to allow the child to progress to the next stage of movement. But in dyslexics, early reflexes may persist. The work of McPhillips et al. also indicates that Dyslexia can be treated by retraining the central nervous system by slowly repeating these movements. Hence the connection between the treatment of Dyslexia and a movement therapy.

The **DysVideo** project at Trinity College was set up to observe the development of 400 children aged below 6 years. Each child is observed through 3 sessions of 20 minutes, each 6 months apart. The session is composed of 14 exercises that are designed to trigger each of four primary reflexes. For example, Fig. 1 shows the movement designed to trigger the ATNR[6, 8] primary reflex. In this movement, the child stands with arms held out in front. The supervisor then turns the subject's head to each side for 5 secs. The arms may follow the head movement or drop. The amount of movement made by the arms gives one clue about the severity of the retained reflex. In a non-dyslexic child, the arms should not move.

MIR'04, October 15–16, 2004, New York, New York, USA.

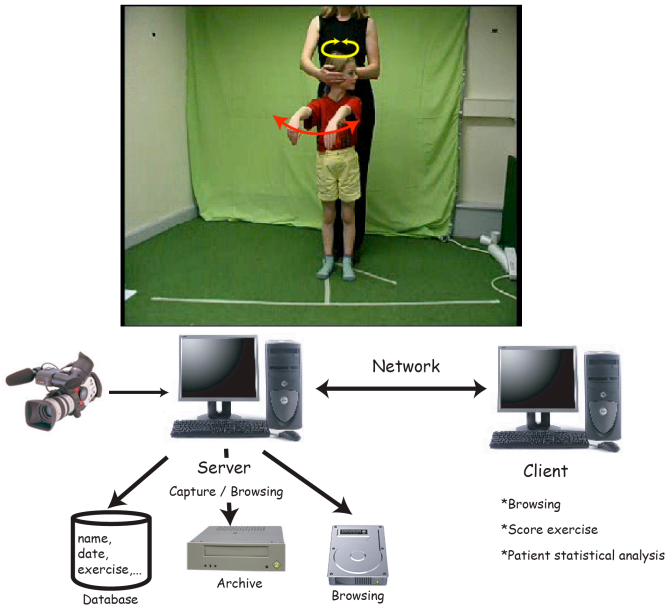

Figure 1: Top: Example of exercise to trigger the ATNR primary reflex. Bottom: Client/Server System Architecture.

The idea is to video each session and then to allow the Psychologists access to the recorded sessions for offline subjective assessment of the degree to which each child meets or fails to meet the required movement template. However, there are clear difficulties that can only be addressed by content-based analysis and indexing as follows.

1. Although sessions may last 20 minutes, the actual measurable information may only be about 5 minutes. This is because much of the time is spent making the child comfortable and setting up each test. Furthermore, children under ten years old are not known for good attention spans, thus intrusive behaviour may cause the session to last even longer. Therefore, it is extremely time demanding for Psychologists to manually locate the useful information from the massive amount of data recorded. *A process is needed to index the start and end of each session automatically.*

2. The movement evaluation as is currently carried out is subjective. Furthermore, without a video record there is no way to cross check retrospectively between different evaluators. Indeed, direct observation requires some training and the movement instance can simply be missed by the observer. Consequently, maintaining a database of scores and movement sessions is essential. *This implies identifying the child and each session uniquely.*

3. Objective movement evaluation is required. This could be achieved by automated tracking of the movement of the limb in question and then attempting to correlate these measurements with a predetermined template motion. However, most trackers require human initialisation. Given the huge database of material within which the usable material is just a fraction, this is impractical. *A mechanism must be found to directly index the active portion of each experiment in order to engage an automated tracker.*

Each of these problems is now addressed in turn.

2. SYSTEM ARCHITECTURE

Fig. 1 shows that the system architecture has a server/client structure. The server performs the capture, indexing and analysis of video sequences, and can also be used as a browser. The different clients browse the captured video sequences remotely. Analysis includes sequence compression and content retrieval.

2.1 Video streaming and compression

A DV camera with an output at a constant bit-rate of 26.4Mbit/s was used. Given that the total video to be stored is about 500hrs; this is equivalent to about 5.8 Terabytes. To keep storage costs low, the DV video stream is compressed using MPEG4 with a 1Mb/s bit-rate. This setting gives comfortable viewing for the human evaluators. The compressed sequences are stored on a disk for "video on demand". Compressed sequences are easily stored in fast access hard drives, e.g. 500hrs of video require 225GB, which is currently easily obtained. Sequences are compressed in real time at the end of the day's recording sessions. For practical (space) and reliability reasons it is more sensible to restrict the recording sessions to one camera only and to avoid streaming direct to disk.

However, 1Mb/s bit-rate does not provide a good enough quality for motion analysis. There are two possible solutions to this problem. (1) The DV media should be processed immediately for motion upon capture. (2) Further attention should be paid to the problem of compressed bit rates required for scientific video analysis. Development of motion analysis techniques is still an active research area and it is not sensible to rely in the future on motion estimates generated once upon capture only. Therefore it is useful to consider the problem of choosing a bitrate which gives little effect on motion estimation, yet yields good enough compression for long term storage. From our experience, the chosen motion estimation process [10] operates properly above a bit-rate of 128Kb/s. Consequently, sessions are compressed at a bit-rate of 2Mb/s, to have a good safety margin, using a MPEG2 codec.

The video from each session is streamed directly into a single file that then must be indexed to indicate the important portion of that file. The system does not create multiple files for each session as it is simpler to maintain a basic database. Thus, key or index files are associated with each session video stream. The creation of the index is discussed below.

2.2 Interactive Audio Markers

The user is asked to use a handheld computer to create tones which are used to indicate the start and end of each exercise (2 digits), as well as the ChildID (6 digits) etc. DTMF tones (Dual Tone Multi Frequency), were used because they are better differentiated from speech and they code 10 digits and two symbols # and *. The symbols are used to

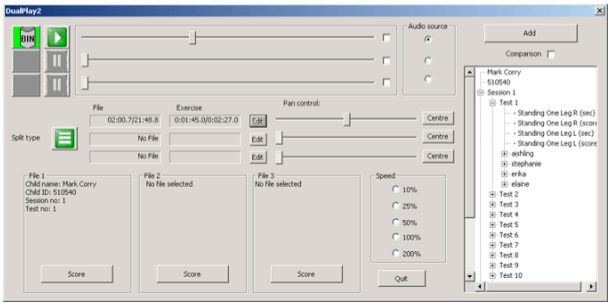

Figure 2: Browser interface. The user can browse and score up to 3 exercises simultaneously.

mark the exercise end or an error, respectively. In the first recorded sessions, the DTMF sound was played, near to the camcorder. Unfortunately, classification was hampered by noise such as laughter. Nevertheless, the detection was successful in more than 95% of the cases.

To achieve 100% accuracy, the DTMF and room audio was recorded on separate channels of the stereo sound camcorder system, thus the detection becomes trivial and 100% accurate. The detection requires the discrimination of two frequencies simultaneously (row/low and column/high) [5]. and consists of 3 steps: 1) measure and threshold of the energy on all DTMF frequencies, 2) identify the key pressed 3) group a set of keys to get the exercise number, the child ID or symbols # and *.

2.3 Browser

The browser allows access to a particular exercise for a given session and child as well as scoring and comparison with other similar sequences. It uses MPEG4 compressed video sequences (1Mbit/sec) and a database (the indexing or key file), which contains time codes to allow random access to particular sessions. The GUI is shown on Fig. 2. It allows the user to watch three different exercises; a window is split horizontally or vertically, when an exercise is added (right-top bottom). Three sliders are used to navigate throughout the exercise, allowing the user to repeatably view the important sections of the session. On the right part of the window, a tree displays information on all exercises taken by a particular child in addition to current user scores. Other user scores can be displayed depending on access rights. The browser also allows computations to derive score statistics across individuals.

3. CONTENT BASED ANALYSIS

Human body analysis is an active research area. A review of methods for identifying and modelling body movement is presented in [12]. Most of the techniques employ multiple camera views in order to facilitate 3D modelling. Problems such as body overlapping, and image clutter are addressed. However, work in vision for human body assessment is focussed on extraction of the entire form while here the interest is in a few objects only.

In contrast with other work, here a single colour camcorder is used. This allows a relatively simple set up, which is essential for non-expert use over long durations. The scene content is almost fully controlled: the background is green, the supervisor is wearing black clothes, marks are placed on the ground to access simple image geometry, and the camcorder orientation is set to have the best view. Colour can therefore be used efficiently for skin detection hence facilitating simple yet robust segmentation.

The fact that a single view is employed, does complicate the analysis. However, by assuming that the orientation of the body in the field of view is roughly the same in each recording, 2D hand tracking yields usable position and motion measurements. The tracking process must be robust in terms of occlusion as well as maintaining lock over long durations of activity. This is discussed later on in this article.

In order to attempt to develop automated motion analysis assessment and explore how well this correlates with the subjective assessment of the psychologists, a mechanism must be found to identify exactly when the exercise actions begin. Efforts are currently concentrated on the ATNR (Asymmetrical Tonic Neck Reflex) exercises. The idea is to use skin detection to locate limbs, and then to use the rough skin information in two ways. Tracking of the centre of gravity (CoG) of the region in the whole frame allows the start of each action to be indexed. Then, closer body localisation can be carried out again using the skin detector. This time the temporal indicators from the CoG analysis can be used to instantiate a tracker for the relevant limb.

3.1 Skin tone detection

Skin detection is a common technique used, e.g., in face recognition [1, 3]. The idea is to associate pixels containing skin with a particular colour distribution that is empirically built from observed images. To detect the skin, we used several colour spaces such as RGB, HSV, rgb, YCbCr and YUV. The best detection quality was obtained using the skin detector described in [7]: a pixel is flagged if "$(R > 95)$ and$(G > 80)$ and$(B > 40)$ and$(R > G)$ and$(R > B)$ and $(R - \min(G, B) > 10)$ and$(R - G > 15)$". This detector avoids selection of pure red or gray pixels. Just before applying this detector, a global colour adjustment is performed to compensate the global colour variations (for unknown reasons, the image becomes randomly blue). Using the carpet colour as the reference colour is the carpet, we simply subtract the colour reference to the colour estimated. A typical result is shown in Fig. 3. In practice all exposed limbs are detected except in instances where the limb colour is changed due to lighting and shadow. Few false alarms also occur in the presence of rich reds. This problem is resolved simply by recommending that subjects do not wear red clothing. The detector works in at least 95% of the cases on 100 sequences of 90s.

4. ANALYSIS OF ATNR (SCHILDER TEST) EXERCISE

This exercise is described in Fig. 1. The aim of the analysis is to track hand positions over the sequence. The analysis is challenging because of many degrees of freedom of arms and hands and unreliable framing of the child in the field of view. Moreover, children do not co-operate actively with the exercise and this implies that less than 50% of the sequences correctly match the exercise template.

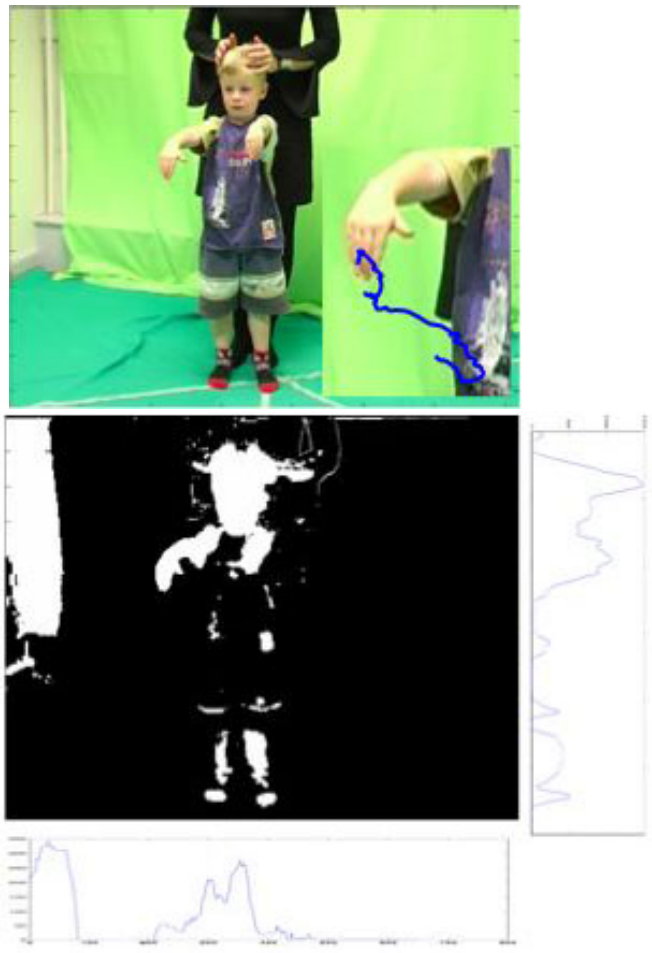

Figure 3: Top: Original with a red indicator showing result of hand detection and an estimated motion track in blue, Bottom: Result of skin detection and horizontal and vertical projections showing body localisation.

4.1 Hand localisation

To build a rough localisation of hands we exploit the starting conditions of the ATNR experiment: hands down, arms up to shoulders and straight forward (e.g. image Fig. 3). Detected areas of skin can then be associated with limbs. Because both arms are detected, a vertical projection of the skin detection image gives the body range along the X axis. Using this narrowed range, a horizontal projection gives the bottom position of the hands with the search constrained in the top half of the image. Fig. 3 shows both projections: we see immediately that it is easy to locate body boundaries.

Once the top part of the body is localised, hands are associated with the lowest parts of that skin/body mass, with small objects removed using an erosion operator with a mask size of 10×10. All possible point pairs p_l and p_r, for left and right hands respectively, are considered. The pair that maximises $p_l(y) + p_r(y)$ is chosen as the hand detected positions.

We tested the hand localisation on two sequences of duration 1.5 mins which is the total extent of each exercise. During the exercise, there are only a few seconds in which the hands are detectable in the expected pose. The hand position is working in 80% of the cases (both hands are correctly localised).

4.2 Analysis

The localisation feature presented above can then be exploited in two ways to provide the Psychologists with a possible objective measure of motion. First of all, there is a need to locate efficiently in time the start and end of each exercise instance. Having done this, hand detection can then be used to initialise a tracker [4] or optic flow field estimation can be used to generate some index of fit to an expected template optic flow field. This paper does not present any results of motion measurement as the study is still in an initial phase. However, the body and hand localisation feature are important features for content access when coupled with simple motion information.

Again exploiting the user context, the ATNR exercise begins with the experimenter's hands moving between head and arms as this is a training period for the child subject. Thus vertical movement is an indicator of the specific start point of this exercise. A simple feature to index this information therefore is the centre of gravity of all the skin detected in a particular frame. This is in fact related to a geometric moment, a feature we have exploited successfully in the past for sport events [4]. A track of the vertical displacement of this Frame-CoG is shown on the right in Fig. 4.

Explicit hand tracking can also yield similar information. Experiments were carried out using a primitive tracker. A hand reference point is assigned which is expected to be at the centre of the palm. Optic flow components within a disc of radius 12 pixels around this reference point are then averaged to estimate the motion into the next frame. In the next frame the point is corrected to be at one half disk radius away from the bottom detected hand portion. Furthermore, to avoid lateral drift, the horizontal position of the reference point is corrected to be at the centre of gravity of the detected hand portion within the disk radius. This correction is at most 3 pixels in practice and hence problems do *not*

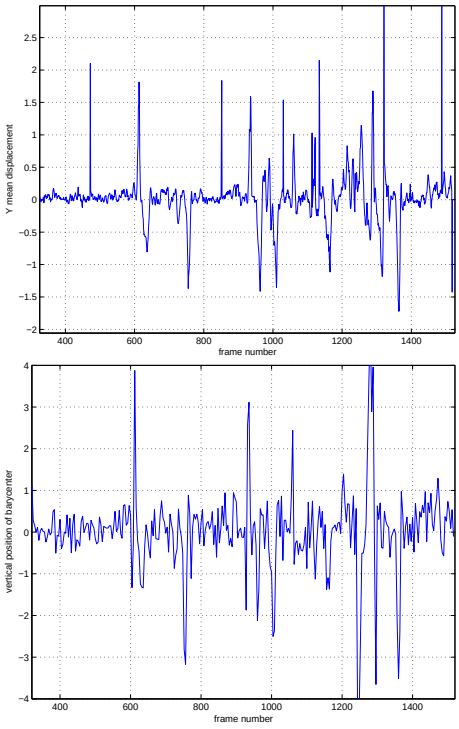

Figure 4: Mean vertical motion of limbs (on the top) and vertical position of the center of gravity (on the bottom)

arise with the hand portion moving outside the disk radius. A track of the hand over 100 frames is shown superimposed in Fig. 3.

Both these features show points of action indexed by large positive motion followed by large negative motion as expected. They do not agree entirely however, since hand tracking is explicit while Frame-CoG is implicit. In practice, we find that using the Frame-CoG feature the start of 90% of ATNR exercises is successfully located, while yielding 25% false alarm rate. The false alarm rate is high due to the crude feature extraction step.

Nevertheless, given a manual initialisation, hand tracking is accurate and over 3 minutes (the full extent of the exercise for two realisations) (as stated in the previous section) there is no loss of lock.

5. ANALYSIS OF ATNR (AYRES TEST) EXERCISE

In this exercise, the child is on all fours, head turned to the camera. The supervisor, seated on one side of the child, turns the head of the child left and right for 5s (see Fig. 5). This movement may trigger a tremor or a bend of the arms. The goal is to measure the angle of the forearm as well as the angle variation and speed for each arm. We have to detect the individual realisation of the exercise since the movement is repeated several times (eyes open and then closed), without inserting marks.

The following process is illustrated in Fig. 5. As in the previous exercise, we apply first the skin detector to select both arms. This selects arms, hair and supervisor hands. Then a bounding box containing both arms is estimated to localise further processing. The bounding box is estimated in two steps. First, a vertical projection gives the vertical position of arms, we search for the two extrema that correspond to individual arms since they are oriented vertically. Second, using the previous vertical boundaries, a horizontal projection is performed. When this projection is scanned from the bottom to the top, this indicates a direction from fingers to upper arms. The maximum of this curve corresponds to hand location since the width of the hand is larger than that of the arm in the view. The vertical extent of the bounding box is taken as three times the hand height. This is a weak hypothesis but valid since since body ratio is relatively constant. The accuracy of the bounding box is 80% on 30 sequences of 90s.

The next step is to estimate the angle of arms and detect each exercise realisation. To do this we fit a line using Andrew's sine robust estimator [2] with sine width set to estimated arm width. This estimator gives a better line fitting than Hough transform or least squares because the arm is not a straight line (due to geometric projection) and because legs and arms are articulated and may merge as a single region. The minimisation implementation is performed using the bisector method by limiting the angle search to $[-\pi/8, \pi/8]$ and origin to $[-w_h, w_h]$ where w_h is the hand width. Line fitting is performed for both arms with the origin set at the corresponding hand location position estimated during the bounding box step.

In Fig. 5 is shown the result of the estimation. The two lower curves represent the angle and the two upper curves are the horizontal position (normalised to fit in the plot) of both hands. From position curves, we can distinguish the actual conduction of the experiment from the preparation stages. The horizontal hand position has to be constant during experimentation since hands are fixed on the ground. Any variations, during a period of few seconds, indicates preparation of the child and not actual conduction of the experiment. Discrimination between preparation and realisation is therefore performed by fixed threshold on the movement curves (called $mc(t)$) : realisation is when $|\text{median}(mc(t), 10) - mc(t)| < 3.5$ where $\text{median}(x, y)$ is the median on x on a window of length y. From the angle curves are extracted the mean, μ_φ, and standard deviation, σ_φ (preparation stage intervals are ignored). Speed is parameterised with mean, $\mu_{\varphi'}$, and standard deviation, $\sigma_{\varphi'}$, on the absolute value of the derivative of the angle (the angle is filtered to reduce the noise by median filter over 20 images). These features are used for motion assessment and are currently under investigation.

The method presented in this section has being ran successfully for 80% of the cases for a set of 30 sequences of 90s each. Failing cases, mainly related to line estimation, are due to bad framing (the child does not fit the image), objects overlapping arms.

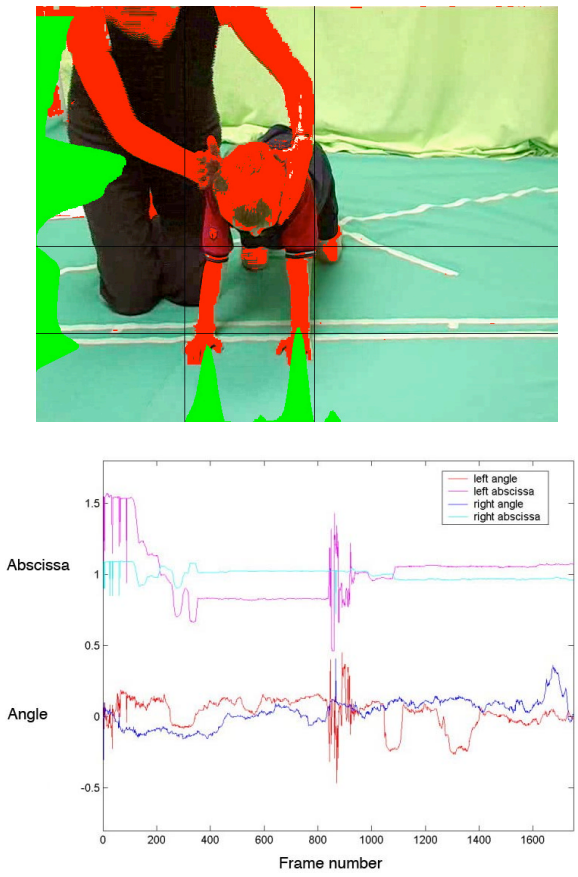

Figure 5: Top: ATNR ayres exercise. Skin detected pixels are in red. The green curves represent the vertical (on the bottom) and horizontal (on the left) projections. Bottom: result of angle and abscissa estimations for both hands.

6. FINAL COMMENTS

This paper has presented a new tool for Psychologists that exploits content retrieval technology in research in motor reflexes in Dyslexia. The system allows video on demand as well as automated indexing and video analysis. As the final users are psychologists and not computer specialists, simplicity and robustness are paramount.

The work has highlighted some interesting implications for massive databases for scientific use. First of all, storage requirements may not enable the best quality material to be stored. This limits the quality of scientific analysis of the picture material. Having two streams of data with two levels of compression appears to be the best compromise. We have presented new results exploring what the breakdown level is for motion accuracy applied to compressed sequences. Secondly, by exploiting the user context, the system is able to deploy 100% reliable indexing. This is imperative for use in scientific investigation. The use of interactive audio cues is novel and allows 100% reliability to be achieved.

New features that yield position information for identifying the start of stylistic movement have also been presented. In this user context, explicit tracking with automated initialisation is possible and this yields powerful information for indexing.

Finally, it is noteworthy that this project has the potential to have a major impact on human observational studies. This project allows for a deep level of data access without the need for 3D observation, by exploiting the user context. Our current work focuses on quantitative evaluation of motion characteristics in dyslexic children. Video sequences showing indexing and parsing output as well as the browser interface are shown at www.mee.tcd.ie/∼sigmedia/dysvideo.

Acknowledgements

This work was funded by Enterprise Ireland.

7. REFERENCES

[1] A. Albiol, L. Torres, and E. Delp. Optimum color spaces for skin detection. *Image Processing, 2001. Proceedings. 2001 International Conference on*, 1:122 – 124, October 2001.

[2] M. Black. *Robust Incremental Optic Flow*. PhD thesis, Yale University, 1992.

[3] D. Chai and K. Ngan. Face segmentation using skin-color map in videophone applications. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(4):551 – 564, June 1999.

[4] H. Denman, N. Rea, and A. Kokaram. Content based analysis for video from snooker broadcasts. *Journal of Computer Vision and Image Undrestanding - Special NUMBER on Video Retrieval and Summarization*, 92(2-3):176–195.

[5] M. Felder, J. Mason, and B. Evans. Efficient dual-tone multifrequency detection using the nonuniform discrete fourier transform. *Signal Processing Letters, IEEE*, 5(7):160 – 163, July 1998.

[6] S. Goddard. *A Teachers Window into a Child's Mind, A non-invasive approach to solving learning and behaviour problems*. Fern Ridge Press, Oregon, 1996.

[7] G. Gomez and E. Morales. Automatic feature construction and a simple rule induction algorithm for skin detection. *Proc. of the ICML Workshop on Machine Learning in Computer Vision*, pages 31–38, 2002.

[8] K. Holt. *Child Development: Diagnosis and Assessment*. Butterworth-Heineman Ltd, 1991.

[9] R. Illingworth. The development of the infant and young child: Normal and abnormal. *Churchill Livingstone, 8th Edition, London*, 1983.

[10] A. Kokaram. *Motion Picture Restoration*. Springer Verlag, 1998.

[11] M. McPhillips, P. Hepper, and G. Mulhern. Effects of replicating primary-reflex movements on specific reading difficulties in children; a randomised double-blind, controlled trial. *Lancet*, (355):537–41, 2000.

[12] C. Sminchisescu. *Three-Dimensional Human Modeling and Motion. Reconstruction in Monocular Video Sequences*. PhD thesis, INRIA, 2004.